Demonstrating ChemGymRL: An Interactive Framework for Reinforcement Learning for Digital Chemistry

Chris Beeler
University of Ottawa
Department of Mathematics and Statistics
Ottawa, Canada
christopher.beeler@uottawa.ca

Sriram Ganapathi Subramanian
University of Waterloo
Waterloo, Canada
Vector Institute for Artificial Intelligence
Toronto, Canada
s2ganapathisubrmanian@uwaterloo.ca

Kyle Sprague
University of Waterloo
Department of Electrical and Computer Engineering,
Waterloo, Canada
kspra093@uottawa.ca

Mark Crowley
Department of Electrical and Computer Engineering,
University of Waterloo,
Waterloo, Canada
mark.crowley@uwaterloo.ca

Colin Bellinger
National Research Council of Canada
Ottawa, Canada, and
Dalhousie University,
Faculty of Computer Science
Halifax, Canada
colin.bellinger@nrc-cnrc.gc.ca

Isaac Tamblyn
Department of Physics, University of Ottawa,
Ottawa, Canada, and
Vector Institute for Artificial Intelligence,
Toronto ON, Canada
isaac.tamblyn@uottawa.ca

Abstract

This tutorial describes a simulated laboratory for making use of reinforcement learning (RL) for chemical discovery. A key advantage of the simulated environment is that it enables RL agents to be trained safely and efficiently. In addition, it offer an excellent test-bed for RL in general, with challenges which are uncommon in existing RL benchmarks. The simulated laboratory, denoted ChemGymRL, is open-source, implemented according to the standard Gymnasium API, and is highly customizable. It supports a series of interconnected virtual chemical *benches* where RL agents can operate and train. Within this tutorial introduce the environment, demonstrate how to train off-the-shelf RL algorithms on the benches, and how to modify the benches by adding additional reactions and other capabilities. In addition, we discuss future directions for ChemGymRL benches and RL for laboratory automation and the discovery of novel synthesis pathways. The software, documentation and tutorials are available here: https://www.chemgymrl.com/.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

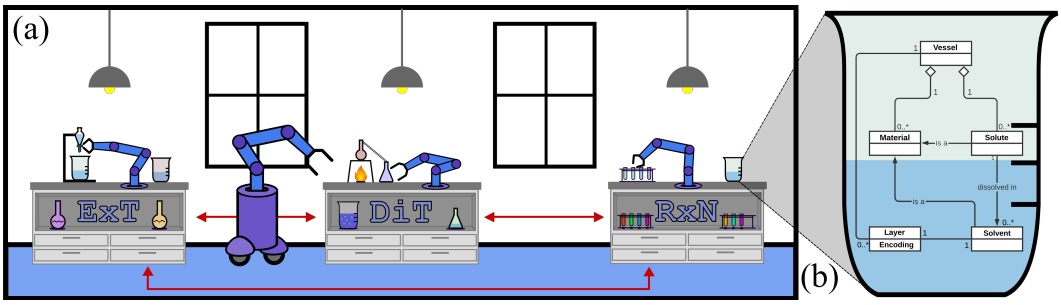

Figure 1: (a) The ChemGymRL environment with individual agents working towards their own goals on the extraction (`ExT`), distillation (`DiT`), and reaction (`RxN`) benches. The user determines which materials the bench agents have access to and what vessels they start with. Vessels can move between benches; the output of one bench becomes an input of another, just as in a real chemistry lab. (b) Materials within a laboratory environment are stored and transported between benches within a vessel. Benches can act on these vessels by combining them, adding materials to them, allowing for a reaction to occur, *etc.*.

1 Introduction

ChemGymRL is a collection of interconnected environments implemented according to the Gymnasium standard [3] for carrying out simulated experiments. As shown in Fig 1, the current version includes reaction, distillation and extraction benches. These provides a mechanism to design, develop, evaluate and refine reinforcement learning (RL) agents for the discovery and optimization of chemical synthesis in a safe and efficient manner. The software, documentation and tutorials can be found at `https://www.chemgymrl.com/`.

RL [11] is class of unsupervised machine learning for sequential decision making, such as determining when and how much reactant to add to achieve a specified objective. RL agents learn by taking actions, making observations, receiving scalar feedback (rewards), and updating a policy. Although, supervised learning is widely adopted in materials research, unlike RL, it requires labeled training data that is often costly or non-existent. As a result, researchers have begun to explore RL for automated chemistry [12, 13, 5]; with ChemGymRL, we aim to expedite progress in this area.

RL's unsupervised nature gives it great potential in AI-guided chemical synthesis and materials discover. With RL, materials researchers can specify a set of requirements or objectives for the new material or synthesis process in advance of its discovery. These can easily be incorporated into the RL reward function, which guides the RL agent to discover novel pathways to target materials in simulation. Moreover, the efficiency of simulated RL experiments enables researchers to define and explore the effect of multiple reward functions. This facilitates the discovery of heterogeneous pathways that trade-off, potentially conflicting, priorities. Thus, researchers apply their depth of knowledge to defining objectives, whilst the RL agent determines the steps to the goal.

The remainder of the paper is organized as follows. Section 2 describes the ChemGymRL environment and primary benches. Section 3 contains a case study on each bench with the Wurtz Reaction, along with pointers to external tutorial resources on installing, setting up and running the benches, defining new reactions, and training off-the-shelf RL algorithms. In addition, we present illustrative results from each bench[1]. Our general conclusions and some ideas for future directions are in Section 4.

[1]Refer to [1] a more detailed set of results and discussions.

2 ChemGymRL

2.1 The Laboratory

The ChemGymRL environment can be thought of as a virtual chemistry laboratory consisting of different benches where a variety of tasks can be completed, see Fig. 1(a) for an overview. The laboratory is comprised of 3 basic elements: **Vessels** contain materials, in pure or mixed form, and track their hidden internal state, **Shelves** are collections of vessels for input/output to benches, **Benches** are simulations of particular chemistry activities.

A bench must be able to receive a set of initial experimental supplies, possibly including vessels, and return the results of the intended experiment, also including modified vessels. Each bench has at least three common components: **Input**: target material given as a one-hot vector, **State**: vessels and contained materials, **Render**: Human Rendering and various possible numeric outputs for learning. In addition, every bench has it's own set of **Actions** and **Rewards**.

2.2 Implemented Benches

Extraction Bench (`ExT`): aims to isolate and extract certain dissolved materials from the input vessels. Actions in ExT include transferring materials between different vessels and utilizing specifically selected solvents to separate materials from each other. The reward is the difference in the relative purity of the desired solute at the first and final steps.

Distillation Bench (`DiT`): aims to isolate certain materials from an input vessel containing multiple materials. The actions are transferring materials between a number of vessels and heating/cooling the vessel to separate materials from each other. The rewards is based on the amount and purity of target in output vessel.

Reaction Bench (`RxN`): allows the agent to transform available reactants into various products via a chemical reaction. The actions include the ability to control the temperature of a vessel and the amounts of reactants added. The reward is calculated after 20 steps have elapsed, the agent receives a reward equal to the molar amount of the target material produced.

Characterization Bench: is used to observe the system. Any observation of a vessel made by an agent must pass through this "bench". Currently, observations occur automatically and are not controlled by the agent. They are visualizations of the ordering of mixture layers, which is similar information as provided in human operated visualization.

3 Case Study

In this section, we use the **Wurtz Reaction**, which are a very commonly used, and well-understood, approach for the formation of certain hydrocarbons to demonstrate ChemGymRL. These reactions are of the form:

$$2\text{R-Cl} + 2\text{Na} \xrightarrow{\text{diethyl ether}} \text{R-R} + 2\text{NaCl}. \tag{1}$$

Wurtz provide an interesting demonstration for this tutorial because the yield varies greatly between each product, making it difficult to train an agent which can optimally make each of them. The tutorial on implementing your own reactions is available in [2].

In the demonstration, we utilize standard RL algorithms implemented with Stable Baselines 3 [9]. The benches ExT and DiT have discrete action-spaces, therefore, Proximal Policy Optimization (PPO) [10], Advantageous Actor-Critic (A2C) [8] and Deep Q-Network (DQN) [7] were used. RxT has a continuous-action space, therefore, Soft Actor-Critic (SAC) [6] and Twin Delayed Deep Deterministic Policy Gradient (TD3) [4] are used in place of DQN.

[2]The tutorial for creating your own reactions is available supplementary material custom_reaction.ipynb.

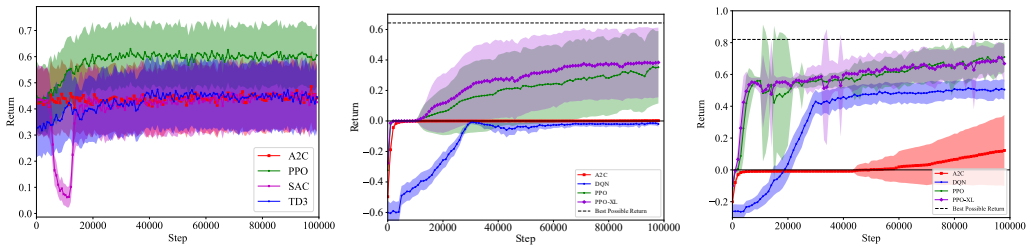

Figure 2: This figure shows average return with $\sigma/5$ shaded, 10 runs for each algorithm with 100K sequential steps x 10 environments. Left: Result for Wurtz `RxN`. Only PPO converges to an optimal policy. Center: Results for Wurtz `ExT`. The PPO and PPO-XL agents consistently acquire positive returns, even approaching the theoretical maximum in some cases. Right: Results for Wurtz `DiT`. The PPO-XL policies outperform the other algorithms both on average and in the best case scenarios.

3.1 Reaction Bench Methodology

For the demonstration of reaction bench (`RxN`), each episode begins with a vessel containing 4 mols of diethyl ether and one of six target Wurtz reaction products involving chlorohexa or sodium chloride. The agent operates for 20 steps and has access to 1.0 mol each of 1, 2, 3-chlorohexane, and sodium to drive the reaction. In addition to SAC, TD3, A2C, PPO, we define a heuristic agent for the experiment, which we expect to be optimal. The tutorial details on setting up and running the Wurtz reaction experiments are provided here[3].

The policy of the heuristic agent is to increase the temperature, and immediately add only the required reactants for the desired product. Following this policy achieves an average return of approximately 0.62. The average return as a function of training steps for each algorithm is shown in Fig. 2 (left). The PPO agents are able to match the heuristic agent for all targets, while some SAC and TD3 agents are able to come close on a few targets. A2C only comes close to the heuristic on producing sodium chloride.

3.2 Extraction Bench

Next, we demonstrate extraction bench (`ExT`) experiment with the target material from the Wurtz extraction. Each episode the agent is given a target material and a vessel containing 4 mols of diethyl ether, 1 mol of dissolved sodium chloride, and 1 mol of one of the 6 Wurtz reaction products. The agent learns to seperate the layers to access the target material.

Since the `ExT` uses discrete actions, we replace SAC and TD3 with DQN. We also use what we call PPO-XL which is PPO trained with more environments in parallel. In addition, we provide a heuristic policy based on what an undergraduate chemist would learn. The tutorial details on setting up and running the extraction experiments are provided here[4].

As seen in Fig. 2 (centre), the agents trained with A2C do not achieve a return above zero, while the agents trained with DQN ended up achieving a negative return. Both PPO and PPO-XL produce agents that achieve significantly more reward than the other algorithms, and outperform the heuristic policy. On average, the best performing agent trained with PPO-XL achieves approximately 10% higher solute purity than the heuristic policy.

3.3 Distillation Bench

Lastly, we demonstrate the distillation bench (`DiT`). Here, we consider a similar experimental set-up to the `ExT` one. Each episode begins with a vessel containing 4 mols of diethyl ether, 1

[3]The tutorial for setting up and running reaction experiments is available in supplementary material reaction_lesson.ipynb.

[4]Tutorial for setting up and running extraction experiments is available in supplementary material extraction_lesson.ipynb.

mol of the dissolved target material, and possibly 1 mol of another material. The agents goal is to maximize the absolute purity of the target material in the vessel.

As before, we have devised a heuristic policy that we expect it to be optimal. The tutorial details on setting up and running the distillation experiments are provided here[5]. In Fig. 2 (right) we can see that on average, the algorithms (excluding A2C) converge faster and with less variation in return than on the other benches. Once gain, PPO and PPO-XL are the best performing agents. They learn a policy that is similar to the heuristic policy to achieving the optimal return.

4 Conclusions and Future Work

Future work will include the development of an agent driven lab manager environment. The lab manager agent will direct and facilitate individual bench agents by deciding which vessel to give to each bench while also specifying the desired goal to each bench. The lab manager agent will make use of the characterization bench, which will have multiple characterization methods with associated costs [2]. The addition of observation costs will improve the agents ability to trade-off competing objectives. The addition of new benches will be explored, allowing more sophisticated experiments to be conducted and new insights into the benefits and challenges of the integration of RL into automated chemistry and self-driving labs.

We have introduced and outlined the ChemGymRL interactive framework with three benches that RL agents can operate and learn in, along with a characterization bench for making observations. We have demonstrated training and assessing standard RL algorithms on each bench. Our results highlight that PPO is generally most effective on each bench, however, there is room for significant improvement in stability and rate of convergence. ChemGymRL is a tool for the community to continue to develop RL for the discovery and optimization of chemical synthesis in a safe and efficient manner.

References

[1] C. Beeler, S. G. Subramanian, K. Sprague, N. Chatti, C. Bellinger, M. Shahen, N. Paquin, M. Baula, A. Dawit, Z. Yang, et al. Chemgymrl: An interactive framework for reinforcement learning for digital chemistry. *arXiv preprint arXiv:2305.14177*, 2023.

[2] C. Bellinger, M. Crowley, and I. Tamblyn. Dynamic observation policies in observation cost-sensitive reinforcement learning. *arXiv preprint arXiv:2307.02620*, 2023.

[3] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[4] S. Fujimoto, H. Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[5] S. K. Gottipati, B. Sattarov, S. Niu, Y. Pathak, H. Wei, S. Liu, S. Blackburn, K. Thomas, C. Coley, J. Tang, et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *International Conference on Machine Learning*, pages 3668–3679. PMLR, 2020.

[6] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. a. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[8] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous Methods for Deep Reinforcement Learning. In *Proceed-*

[5] Tutorial for setting up and running distillation experiments is available in supplementary material distillation_lesson.ipynb.

ings of The 33rd International Conference on Machine Learning (ICML), pages 1928–1937, 2016.

[9] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *The Journal of Machine Learning Research*, 22(1):12348–12355, 2021.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[11] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*, volume 9. MIT Press, Cambridge, MA, 1998.

[12] A. A. Volk, R. W. Epps, D. T. Yonemoto, B. S. Masters, F. N. Castellano, K. G. Reyes, and M. Abolhasani. Alphaflow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nature Communications*, 14(1):1403, 2023.

[13] Z. Zhou, X. Li, and R. N. Zare. Optimizing chemical reactions with deep reinforcement learning. *ACS central science*, 3(12):1337–1344, 2017.